Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 138 (2026) 90–95

www.elsevier.com/locate/procedia

18th CIRP Conference on Intelligent Computation in Manufacturing Engineering

Mobile, multimodal, vision-based data acquisition system for passive monitoring in production and intralogistics

Keno Moenck[a,*], Philipp Prünte[a], Jonathan Determann[a], Eidan Erlich[a,b], Dhananjay Patki[a,b], Frank Bitte[c], Martin Gomse[a], Thorsten Schüppstuhl[a]

[a]*Hamburg University of Technology, Institute of Aircraft Production Technology, Denickestraße 17, 21073 Hamburg, Germany*
[b]*University of Waterloo, 200 University Ave W, Waterloo, ON N2L 3G1, Canada*
[c]*Airbus Operations GmbH, Kreetslag 10, 21129 Hamburg, Germany*

* Corresponding author. Tel.: +49-40-42878-3341; fax: +49-40-42731-4551. *E-mail address:* keno.moenck@tuhh.de

Abstract

The digitalization of chaotic intralogistics and production processes, including, e.g., humans and otherwise dynamic or static, non-tracked assets, as in the case of lot size one and large-object production facilities, requires non-invasive sensor solutions. One approach is to equip already movable assets on the shopfloor with multimodal 2D/2.5D/3D optical sensor systems that perceive the surrounding environment – such a solution requires methods for sensor calibration, sensor fusion, localization, and mapping. Besides, to comply with data privacy regulations, data must be de-personalized. This work proposes a mobile, multimodal sensor system that passively monitors the surroundings, localizes itself, outputs de-personalized data online, and can recreate the environment as a geometric digital twin.

Peer-review under responsibility of the scientific committee of the 18th CIRP Conference on Intelligent Computation in Manufacturing Engineering (CIRP ICME ‘24)

Keywords: monitoring; vision-based perception; manufacturing optimization; factory planning

1. Introduction

Monitoring intralogistics, production, and the enclosing surroundings on the shopfloor can serve optimization and control of the products, processes, and resources. In non-deterministic settings, sensor-based solutions are specifically of interest since the chaotic characteristic induced through, e.g., non-automated/-controlled assets, demands the acquisition of in-situ data. For example, an aircraft as a variant-rich and large product is assembled at a low cycle time with a fixed position characteristic, where resources move to the product, often manually, whereas load carriers, parts, and other assets might be positioned unknown [1, 2].

Sensor systems that can perceive parts of a shopfloor’s current state can work inside-out or outside-in. The latter refers to a system that perceives the object of interest from the outside, e.g., tracks a resource’s position wirelessly through fixed-positioned anchors and moving tags [3, 4]. The disadvantage is that the overall observation space must be equipped with a dense net of anchors, and each to-be-tracked asset must also be tagged. An inside-out system perceives the environment from its point of view, e.g., an Automated Guided Vehicle (AGV) localizes itself by sensing for defined magnetic points in its environment [5].

Wanting to not only perceive, e.g., taggable assets like AGVs or other load carriers but also free-movable “things” like pallets, humans, or jigs, a vision-based solution provides much denser information. An outside-in system demands several vision sensors, for which numbers increase with decreasing occluded area, while an inside-out system on existing moving

2212-8271
Peer-review under responsibility of the scientific committee of the 18th CIRP Conference on Intelligent Computation in Manufacturing Engineering (CIRP ICME ‘24)
10.1016/j.procir.2026.01.017

vehicles would, however, only perceive things from the prominent movement areas. Both such systems can complement each other; however, we argue that, especially in fixed assembly line production of large-sized components, the latter approach suits moving assets that travel through various warehouses and shopfloors.

The information density given by vision systems is high, such that privacy data, e.g., the movement of operating personnel or even their identification, is exposed within aforesaid data-hungry systems. Here, privacy regulations demand the de-personification of the data, so privacy is not publicized throughout the whole control and monitoring chain.

Therefore, in this work, we elaborate on a mobile, multimodal, vision-based inside-out data acquisition system that can be utilized to monitor various assets on the shopfloor passively; we contribute threefold:

- We propose an exemplary sensor system equipped with 2D, 2.5D, and 3D modalities to perceive the environment and develop a localization and mapping strategy to allow the sensor system to localize itself inside an unknown environment.
- We present an approach to de-personalize the 2D image data by masking humans.
- We present a pipeline to reconstruct the environment from the sensor data three-dimensionally, representing a geometric digital twin of the shopfloor that can serve further monitoring and control applications.

The rest of this work is structured as follows: First, in the related works section (s. Section 2), we outline the essential system components and introduce state-of-the-art approaches to preserving privacy in vision data. We introduce the data acquisition system in Section 3 and briefly describe the implementation of our test setup in Section 4. In Section 5, we present our experiments and summarize as well as conclude this work in Section 6.

2. Related Works

2.1. System Components of a Mobile Sensor System

A multimodal mobile sensor system, as in the given context, must have four essential core components: a variety of input sensors, a localization system, a framework for networking and communication, and an information component to provision the data online and/or offline. We discuss the individual possible sensors that serve input data in Section 3. In the following, we outline approaches to localization and the information processing component providing networking, communication, and data provisioning.

Localization typically comes with the additional question for mapping. Approaches that solve the problem of localization while concurrently building a map are termed Simultaneously Localization and Mapping (SLAM). Localization solely either needs an existing map of the surroundings or estimates odometry-like movements between a few consecutive frames. Without any prior map, SLAM is mostly the choice given computationally capable hardware; frame-to-frame estimation with dead reckoning is too prone to drift [6].

The information brick in robotic applications is typically abstracted by a robot middleware that lies between the operating system and modular software blocks. Since existing frameworks perfectly adapt through abstraction to heterogenous actuators, sensors, and software [7], they perfectly fit even applications without direct actuation, as in our case. Such a robust framework for networking, communication, and data transport is the Robot Operating System (ROS) [8]. ROS facilitates node-based uni- and bi-directional communication as well as features for managing transformations, time, and efficient data transfer. All timestamped message streams can be saved as files, which can be used to replay the data during offline processing. ROS offers online and offline sensor data processing capabilities while providing the necessary framework components for networking, communication, and data provisioning.

2.2. Preserving Privacy in Vision Data

To date, the majority of methods used for preserving privacy in vision data rely on some form of image segmentation and obfuscation. First, areas of concern are identified, and then they are manipulated through, e.g., pixelation, blurring, or zeroing out entirely. While traditional/rule-based segmentation is possible, most contemporary methods use machine learning-based algorithms to more robustly identify areas or objects of concern with higher generalization in dynamic scenes.

Chen et al. demonstrate the use of GANs to infill the problematic areas with representative AI-generated data [9]. Kim et al. use a CNN-based approach to perform video infilling [10], while Li. et al. augment this approach with optical flow methods [11]. Further, Lee et al. use a “copy-and-paste” method to infill, using information in surrounding frames to fill in areas in the current frame [12]. Generative methods have the potential to hallucinate in objects that were not initially included in the scene, while the copy-and-paste method only uses information from other frames present in the recording. In the scope of this work, we argue that generatively inpainting the problematic areas is not that of interest, while masking is sufficient, as generating content does not provide any additional information.

3. Vision-based Data Acquisition System

The following section elaborates on the overall system setup and the individual functional components.

3.1. System Overview

Figure 1 depicts the mobile, multimodal sensor system with inside-out perception capabilities. On the sensor side, we chose a variety of 2D cameras, 2.5D sensors, and a 3D LiDAR. The

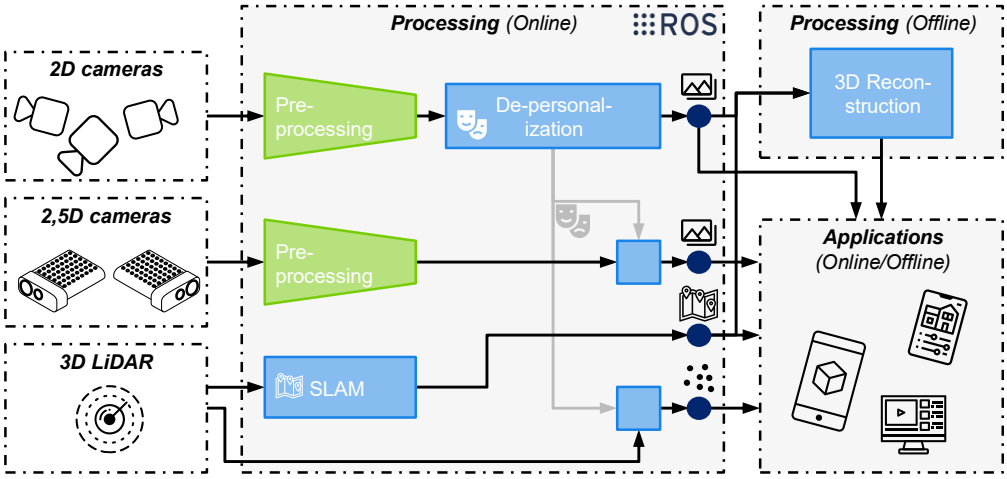

Figure 1: Overview of the mobile, multimodal sensor system concept.

widest Field of View (FoV) is typically the one of the LiDAR but is often constrained to only a few layers, impairing the vision in the vertical direction. However, LiDARs have a much higher sampling rate than cameras, are distance-aware, and thus typically preferred to be used in SLAM instead of utilizing visual odometry-assisted SLAM solely. We, therefore, intend the 3D LiDAR for the sensor system's localization and provide the map for online applications, but for offline applications, we will derive a more fine-grained 3D reconstruction based on Structure-from-Motion (SfM) later. The SLAM's output, on the one hand, tags the other acquired data with localization information in a generated map; on the other hand, the map can be later used in offline applications.

The 2D/2.5D sensor modalities are first pre-processed. During this step, the system applies basic image cropping and color correction. The de-personalization uses solely the 2D image data because of its information-richness compared to the 2.5D and 3D modalities. However, the generated masks can then be projected onto the other modalities through the intrinsic and extrinsic information.

All the mentioned processing steps are done online. All outputs are provided timestamped to processing done offline or for further applications.

3.2. Sensor Calibration

We performed camera calibration using a ChArUco calibration board. After capturing images that include the board from various angles, we applied standard calibration to obtain the intrinsic matrices for each camera. Overlapping Fields of View (FoV) are required for extrinsic calibration. We took images of the ChArUco board so that a significant part of the board was visible in at least two of the cameras. We then applied the SolvePNP [13] method to solve for the transformations between the cameras.

Since the 2.5D sensors we use provide extrinsic between the RGB and depth image, we automatically achieved extrinsic calibration to all other cameras.

3.3. Lidar-based Localization and Mapping

We chose Kiss-ICP [14] as the SLAM method, which performs comparatively on benchmark datasets, like KITTI [15], and is robust against changing environments. It employs Iterative Closest Point (ICP) to discern correlations within LiDAR point clouds, subsequently enabling sensor localization. Four core features characterize Kiss-ICP: It utilizes a simplified dynamic model and a scan deskewing process to enhance data quality. Moreover, it incorporates spatial downsampling of point clouds and develops a local map to estimate incremental poses and point correspondences. An adaptive threshold nearest neighbor search is then employed to identify corresponding points between each point cloud. Subsequently, it optimizes point-to-point ICP using the predictive model and point correspondences before integrating the points into the local map.

3.4. De-personalization

We perform de-personalization using the YOLOv8 [16] model to segment frames from each camera's video feed into areas that contain humans and areas that do not. Areas containing humans are either modified by masking or applying a median blur. Areas without humans are kept identical to the original frame. The masks and the masked images are passed to further processing steps.

In order to minimize the computational requirements of segmentation, the number of classes is limited to one (humans). Furthermore, the threshold for human detection was lowered to 20% confidence in order to detect humans more consistently. To uphold data protection and privacy standards, minimizing false negatives is imperative to the system's reliability. False negatives may lead to undesirable outcomes, such as legal liability and consequences, particularly if the data is misused or exploited. As such, the increase in false positives is considered a worthwhile trade-off.

3.5. Reconstruction

The offline processing step reconstruction iterates through the pre-processed raw images. First, histogram equalization is applied to the images to maximize the contrast. Then, gamma and illumination corrections are applied to standardize the illumination and color space across the images, maximizing the features in the images and ensuring that reconstruction will occur in the same color space. Subsequently, SIFT [17] features are derived from the images to detect keypoints and generate descriptions. Each image is assigned a similarity score relative to all others. Images lacking sufficient information are excluded, while those within a specified similarity threshold are retained, resulting in approximately a halving of the dataset size in a typical capturing run, given a frame recording rate of 10Hz. SIFT is preferred for image matching due to its scale and orientation invariance, which is critical when comparing images captured from diverse camera angles and sizes. SIFT operates by establishing a scale space where features remain scale-independent. Keypoints are then localized within the image, followed by orientation normalization to mitigate rotational discrepancies. Unique descriptors are assigned to each keypoint for subsequent image matching.

Our reconstruction pipeline utilizes COLMAP [18], which is a Multi-View Stereo vision (MVS) [19] and SfM pipeline. SfM entails reconstructing a 3D scene through motion parallax. COLMAP has two primary phases. First, a correspondence search is done to identify overlapping areas between images. For each image, keypoints are extracted and matched with others using SIFT, followed by geometric verification of the overlapping regions. Images are deemed geometrically verified if a specific number of points pass validation. The subsequent reconstruction is incremental. Bundle adjustment is finally employed to rectify drift resulting from image registration and triangulation, minimizing projection error.

We utilize the localization priors for each image to reduce the time for neighbor search and diminish keypoints in masks generated during the de-personalization, resulting in humans not being part of the 3D reconstruction.

4. Implementations

The current test setup is depicted in Figure 2. The sensor array includes 3x IDS uEye 5280-CP-C-HQ RGB cameras, 2x Microsoft Azure Kinects, and 1x SICK MRS 1104C-111011 LiDAR. Online processing is conducted on an Intel NUC with an Intel i7-1165G7 processor, 64 GB of RAM, and an NVIDIA RTX 2060 GPU located on the mobile sensor system. The test setup has a 750Wh battery to allow mobile operations. The offline post-processing was conducted on an AMD EPYC 7443P CPU, 2x NVIDIA RTX A6000 GPUs, and 128 GB of RAM.

Figure 2: Sensor setup used during the experiments.

5. Experiments

This section presents qualitative studies on de-personalization and quantitative studies on 3D reconstruction. The area under test was our research laboratory, as shown in Figure 3. Figure 4 depicts samples from a test run in the laboratory, including the images from the 2D camera rig setup and the map that KissICP generates. The top three cameras' FoVs only slightly overlap, so we could assemble a panoramic image while increasing the total FoV. Since we positioned the Kinects in forward orientation such that we have 2D and 2.5D sensor alignment, the Kinect's 2D images face a similar FoV as the IDS cameras.

Figure 3: 3D scanned ground truth data of our research laboratory captured with a Leica BLK360G2.

5.1. Quantitative Results

We run the 3D reconstruction pipeline on a test run from the laboratory, which results in a point cloud of 10,497k points. A top-view projection is depicted in Figure 5. We applied basic Statistical Outlier Removal (SOR) with $d_{max} = \mu + 1.0 \cdot \sigma$ and $k = 20$, as well as voxel downsampling to $d = 0.05$. After that, we ended up with 836k points, which is an amount manageable on a consumer hardware device.

If we look at the reconstruction, we observe a variety of artifacts, especially at objects' borders; these are typical since images are not perfectly aligned, intrinsics are not perfect, and distortions are not exactly modeled during reconstruction. It depends on the application of how these influence further processing since, with standard filter techniques like SOR, we can diminish them only to a certain extent.

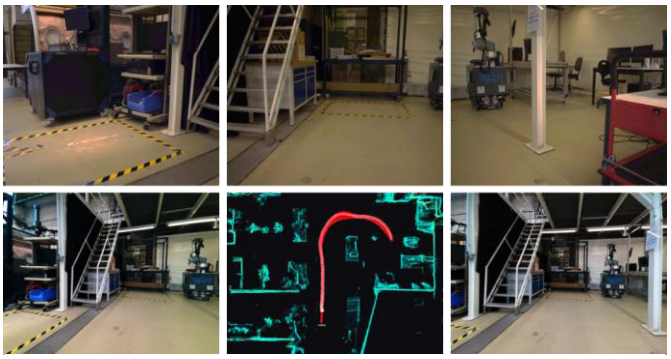

Figure 4: Sampled images from our sensor array (top row: IDS cameras; bottom row left and right: Kinect; bottom middle: position at sampling time and the generated map).

To get an insight into how well the 3D reconstruction accords with the ground truth, we manually aligned them while fine-tuning the transformation using ICP. Subsequently, we derived the point-to-point distances without local modeling. Considering a normal distribution, 68.27% of the points have less distance to the ground truth than 11.64 cm. Taking into account that the reconstructed point cloud is voxel downsampled, aligned in post-processing, and the ground truth is sampled from different viewpoints, we think the data's resulting uncertainties are sufficient for various real-world applications, including, e.g., factory planning, optimization, or retrofit. We additionally split the cloud into points outside $1 \times \sigma$, which are depicted in Figure 6.

Figure 5: 3D reconstruction of the laboratory environment with our Structure-from-Motion (SfM) pipeline.

The remaining points are especially in areas that were far away from the cameras during the test run. This leads us to the point that the measurement data will have less uncertainty in the specific areas of movement than in the far-distance regions that are probably not that of interest. We further suggest that the point cloud could be cropped close to the navigation path in a post-processing step.

5.2. Qualitative Results

Given the mentioned uncertainties, the quality of the 3D reconstructed scene through SfM suffices applications that demand perceiving larger-scaled objects. Considering the de-personalization, we observed that in uncrowded scenes, the reconstruction still retains its performance, while given crowded scenes, the SfM pipeline will mostly fail. Since we do not in-paint the critical image parts, false hallucinations or blurred areas are not part of the reconstruction.

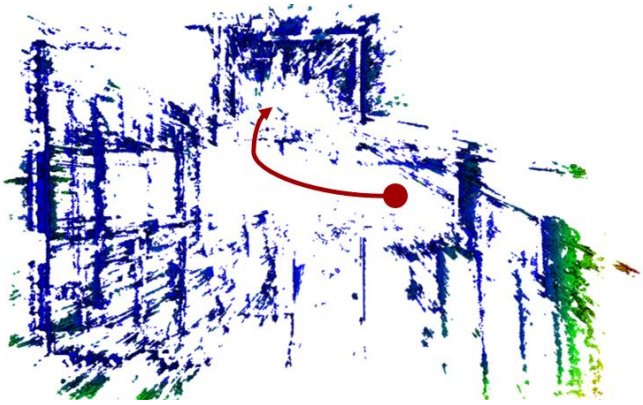

Figure 6: Points of the reconstructed scene outside $1 \times \sigma$ and the approximate path during acquisition (blue colored points have less deviation than green < yellow < red).

Figure 7: De-personalization of humans in images.

Figure 7 depicts the results in two samples that show qualitatively the performance of the de-personalization method. In the image on the left side, a human sits on a chair in front of a desk. Parts of the legs are false positives; however, the critical parts, like the upper body and the head, are masked. Further, in the image on the right side, we observe that the YOLO model also performs well in images with motion blur, which suffices applications where the sensor system moves slowly. Given faster motion speeds, we either have to decrease the shutter time or hope that the motion blur already makes identifying critical parts of a human hard.

6. Conclusion and Outlook

In this work, we elaborated on a mobile, multimodal, vision-based data acquisition system that can either utilize already existing sensor systems on a moving asset or can be non-invasively retrofitted. The integration of 2D, 2.5D, and 3D sensors provides a robust mechanism for capturing detailed environmental data without disrupting ongoing operations. Online digitalization of production and intralogistics is

especially interesting when facing non-deterministic processes with a particular uncertainty, e.g., given a lot of manual actions from assembly to intralogistics, as in the aircraft industries' producing activities. Besides, the given 3D reconstruction pipeline supports the digital recreation of the environment representing a geometric digital twin, which can, e.g., serve factory planning activities. Subsequent applications can serve, e.g., change detection of assets (based on 2D/2.5D data) or resource optimization through analyzing available moving spaces in 3D. In concurrent work, we elaborate on additional online applications based on the proposed concept [20]. The experimental studies underscore the proposed system's potential to offer online and offline insights into intralogistics and production operations. The de-personalization process implemented within the system ensures that privacy concerns are adequately addressed, maintaining the integrity of personal data while still leveraging the benefits of dense, online-acquired environmental data.

Looking forward, the scalability of the presented system across different contexts and its integration with existing digital frameworks remain a key question for further research. Besides, future work will focus on refining applications and use cases as well as applying novel techniques to the 3D reconstruction pipeline, e.g., NERFs, a technique for novel view synthesis, which would bring the shopfloor street view-like to the topfloor.

Acknowledgments

This work is part of the research project *Produktionssysteme der digitalisierten Luftfahrtindustrie auf Basis effizienter Service-Architekturen (ProDigieS)* under the grant number 20D2123E, supported by the *Federal Ministry for Economic Affairs and Climate Action (BMWK)* as part of the *Federal Aeronautical Research Programme LuFo VI-2*.

CRediT author statement

Conceptualization: K.M.*, P.P.*, J.D.*, E.E., D.P., F.B.; Software: E.E.*, D.P*; Validation: E.E.*, D.P.*, P.P., J.D., K.M.; Formal analysis: E.E.*, D.P.*, K.M.; Investigation: E.E.*, D.P.*, P.P., J.J., K.M.; Resources: M.G., T.S.; Data Curation: E.E.*, D.P.*, P.P., J.D., K.M.; Writing - Original Draft: K.M., E.E., D.P.; Writing - Review & Editing: K.M., P.P., J.D., F.B., M.G., T.S.; Visualization: K.M., E.E., D.P.; Supervision: K.M., P.P., J.D., F.B., M.G., T.S.; Project administration: M.G., T.S.; Funding acquisition: M.G., T.S. All authors have read and agreed to the published version of the manuscript; * equal contributions and leads. The work from E.E. and D.P. was done during their internship at the Institute of Aircraft Production Technology.

7. References

[1] K. Moenck, J.-E. Rath, J. Koch, A. Wendt, *et al.,* 2023. Digital twins in aircraft production: Challenges and opportunities. doi:10.13140/RG.2.2.15698.53441/1.

[2] F. Gehlhoff, H. Nabizada, M. Weigand, L. Beers, *et al.*, 2022. Challenges in automated commercial aircraft production. doi:10.1016/j.ifacol.2022.04.219.

[3] C. Hegedus, A. Franko, P. Varga, 2019. Asset and production tracking through value chains for industry 4.0 using the arrowhead framework. doi:10.1109/ICPHYS.2019.8780381.

[4] Rácz-Szabó, A., Ruppert, T., Bántay, L., Löcklin, A. *et al.*, 2020. Real-Time Locating System in Production Management. doi:10.3390/s20236766.

[5] Bourny, V., Capitaine, T., Barrandon, L., Pegard, C. *et al.*, 2010. A localization system based on buried magnets and dead reckoning for mobile robots. doi:10.1109/ISIE.2010.5637693.

[6] Yousif, K., Bab-Hadiashar, A., Hoseinnezhad, R., 2015. An overview to visual odometry and visual slam: Applications to mobile robotics. doi:10.1007/s40903-015-0032-7.

[7] Elkady, A., Sobh, T., 2012. Robotics middleware: A comprehensive literature survey and attribute-based bibliography. doi:10.1155/2012/959013.

[8] Robotic Operating System. https://www.ros.org.

[9] Chen, Z., Zhu, T., Xiong, P., Wang, C. *et al.*, 2021. Privacy preservation for image data: A gan–based method. doi:10.1002/int.22356.

[10] Kim, D., Woo, S., Lee, J.-Y., Kweon, S., Deep video inpainting. doi:10.48550/arXiv.1905.01639.

[11] Li, Z., Lu, C.-Z., Qin, J., Guo, C.-L., Cheng, M.-M., 2022. Towards an end-to-end framework for flow-guided video inpainting. doi:10.48550/arXiv.2204.02663.

[12] Lee, S., Oh, S.W., Won, D., Kim, S.J., 2019. Copy-and-paste networks for deep video inpainting. doi:10.1109/ICCV.2019.00451.

[13] Marchand, E., Uchiyama, H., Spindler, F., 2016. Pose estimation for augmented reality: A hands-on survey. doi:10.1109/TVCG.2015.2513408.

[14] Vizzo, I., Guadagnino, T., Mersch, B., Wiesmann, L. *et al.*, 2023. Kiss-icp: In defense of point-to-point icp – simple, accurate, and robust registration if done the right way. doi:10.1109/LRA.2023.3236571.

[15] Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. doi:10.1109/CVPR.2012.6248074.

[16] Ultralytics YOLO, 2023. https://github.com/ultralytics/ultralytics.

[17] Lowe, D.G., 2004. Distinctive Image Features from Scale-Invariant Keypoints. doi:10.1023/B:VISI.0000029664.99615.94.

[18] Schonberger, J.L., Frahm, J.-M., 2016. Structure-from-Motion Revisited. doi:10.1109/CVPR.2016.445.

[19] Schönberger, J.L., Zheng, E., Frahm, J.-M., Pollefeys, M., 2016. Pixelwise View Selection for Unstructured Multi-View Stereo. doi:10.1007/978-3-319-46487-9_31.

[20] Prünte, P., Determann, J., Moenck, K., Erlich, E., Patki, D., Bitte, F., Gomse, M., Schüppstuhl, T., 2024. Leveraging passive monitoring applications in production and intralogistics